Macro- and Micro-rhythm in L2 English: Exploration and Refinement of Measures

Ortega-Llebaria, M, Silva Jr, L, Nagao, J.
University of Pittsburgh, State University of Paraiba, Gifu Shotoku Gakuen University
mao61@pitt.edu, leonidas.silvajr@servidor.uepb.edu.br, junnagao@gifu.shotoku.ac.jp

ABSTRACT

This paper constitutes a pioneering attempt to explore the pitch-based macro-rhythm measures proposed by Jun [4], namely macroR_Var and macroR_Freq, in a large L2-speech database and relate them to micro-rhythm measures based on vowel and consonant interval durations. It also proposes a refinement of Jun's macroR_Freq. Eleven 2.5-minute TED talks in L1 English and their corresponding imitations by 11 Japanese English-learners were analyzed using semi-automated labeling and scripts. After intensive practice, Japanese speakers successfully imitated pitch contours, reaching target-like macroR_Var scores while failing at imitating macroR_Freq scores and the duration variation of vowel and consonant intervals as illustrated by several Varco and nPVI micro-rhythm measures. Relevantly, the newly proposed macroR_Freq not only was sensitive to L1-L2 speech differences but was also modulated by several micro-rhythm measures, illustrating for the first-time the nature of the micro-rhythm and macro-rhythm interaction in the Japanese L1-English L2 interface. Possible cognitive and linguistic based explanations are discussed.

Keywords:
L2 rhythm, Macro-rhythm, L2 acquisition, L2 Prosody, EFL

1. INTRODUCTION

Research in L2 rhythm has been largely based on duration-based micro-rhythm measures (microR), e.g., VarcoV, VarcoC, nPVI_V [1], [2], which compute duration variations of the vowel and consonant intervals within an utterance. These measures capture cross-linguistic differences in syllabic structure and the expression of word prominence, which differ between traditional rhythm-classes (for a critical review, see [3] and [4]). For example, stress-timed languages like English –where stressed syllables have long full vowels, unstressed syllables have short reduced vowels, and syllables have complex onsets and codas – tend to have more duration variation of vowel and consonant intervals than syllable-timed languages like Spanish and mora-timed languages like Japanese – which have simpler syllable structures and no vowel reduction. As a result, English obtains lower %V, and higher VarcoV and VarcoC scores than Japanese and Spanish [5].

A newer set of rhythm measures based on the repetition of F0 patterns was proposed by Jun [6]. These measures, referred to as macro rhythm (macroR), capture patterns of F0 repetition above the word level. For example, macroR computes the distance between F0 peaks and F0 valleys within an utterance. So far, macro-rhythm measures have been tested across different L1 and speaking styles, e.g., [7], [8], but, as far as we know, not yet in L2.

While there is a body of research on the imitation of L1 sentence prosody (cf. [17], [18] among others), to our knowledge, this paper is the first attempt to explore macroR measures in a large L2-speech database of Japanese learners of English. The L1 Japanese – L2 English context is especially interesting as the two languages largely differ in their microR and macroR strengths. Japanese[1] has a weaker microR than English since not all content words in Japanese have pitch accents while all content words in English have stress which is cued not only by pitch but also by duration. With regards to macroR, it is predicted to be stronger in Japanese than English [6] because English has a larger inventory of pitch-accents than Japanese increasing variation. In contrast, Accentual Phrases in Japanese have a fixed hat-shaped pitch-contour that repeats along the utterance. Thus, Japanese-accented English is expected to display a weaker microR and a stronger macroR than L1 English.

To analyze this large amount of L2-speech data we used webMAUS forced-aligner [9] and two scripts for Praat [10], [11] that automatically post-processed the alignment into new phonetic units, corrected F0 tracking errors, and computed macroR and microR measures, and pauses. Moreover, a new macroR frequency index, based on the ratio of the F0 peaks and F0 variance (macroR_Freq_F0var), is proposed (see 2.3.1.).

The research questions we are addressing are the following:
[1] Are microR measures sensitive to Japanese accented English as to convey the expected weaker microR in comparison to L1 English?
[2] Are macroR measures sensitive to Japanese accented English as to convey the expected stronger macroR in comprison to L1 English?
[3] Is there any relation between macroR measures and microR measures in Japanese-accented English?

2. METHODOLOGY

2.1. Participants and Task.

Eleven Japanese students of English enrolled at Gifu Shotoku Gakuen University, Japan, participated in this study. They were 18-21 years old, 4 males and 7 females. They participated in English Recitation Contest where they recite a 2.5-minute TED talk or English speech of their choice in front of a team of judges and a large audience. They rehearsed for an average of 2 months with feedback from their teachers and peers to faithfully reproduce the content and emotion of the original talk. Altogether, the 11 imitations along with the corresponding original TED talks resulted in a 55-minute speech sample (31-minute L2 English, and 26-minute L1 English).

2.2. Data processing

After transcribing the original TED talks, the students' imitations, and editing recordings for noise (e.g., cutting off applauses), transcriptions and sound files were input into the aligner to obtain TextGrids which segmented words into phonemes. When needed, TextGrid labels were manually adjusted. Following protocols in [19], each TED talk was divided into 10 to 15 chunks, i.e., utterances between pauses that conveyed a coherent meaning and ended with a final (non-continuant) boundary tone. These TextGrids were input into the scripts VVUnitAligner and SpeechRhythmExtractor which output for each chunk the measures described in the section below.

2.3. Measures

2.3.1. Micro-Rhythm Measures

Micro-Rhythm (microR) in this paper refers to the measures based on the duration of the consonant and the vowel intervals within a chunk, e.g., percV calculates the proportion of vowel intervals within each chunk. As explained in the Introduction, these duration-based measures capture cross linguistic differences in syllable structure and word prominence, e.g., in languages with simpler syllabic structures like Japanese, percV scores around 50%, while in languages with complex onset and codas like English values below 50% are expected. The variation coefficient of vowels and consonants (VarcoV and VarcoC respectively) are calculated by the ratio of the standard deviation of vowel or consonant intervals and the interval mean value. The higher the Varco, the more variable the duration of the analyzed intervals [2]. This variation tends to be higher in languages like English, where lexical stress is cued by duration, in comparison to languages like Japanese, where lexical pitch-accents are not cued by duration. The Normalized Pairwise Variability Index for vowels and consonants (nPVI-V and nPVI-C respectively) is computed by the ratio of the differences between consecutive vowel or consonant intervals and the sum of the duration of the respective intervals [1].

2.3.2. Macro-Rhythm Measures

Jun [6] defined Macro-Rhythm (macroR) as the phrase-medial tonal rhythm conveyed by the repetition of F0 patterns. The more regular this repetition is, the stronger macroR becomes. She proposed two macroR measures to capture it, namely, macroR_Var as in (1) and macroR_Freq as in (2).

$$(1)\quad MacR\_Var = rSD + fSD + pSD + vSD$$

$$(2)\quad MacroR\_Freq = \frac{number\ of\ F0\ peaks}{number\ of\ Prosodic\ Words}$$

MacroR_Var captures the F0 contour of a chunk. Larger scores in (1) come from larger variations of rising (rSD) and falling (fSD) F0 excursions as well as larger duration variations between F0 peak-to-peak (pSD) and valley-to-valley (vSD). Thus a higher score in MacroR_Var indexes more variation, and therefore, weaker macroR. MacroR_Freq captures the synchronization of the F0 peaks in an utterance with prosodic words. A score close to one in (2) comes from having one F0 peak per each Prosodic Word conveying a strong macroR. Thus, while both measures capture macroR, macroR_Var focuses on the chunk F0 contour without reference to the segmental content that carries it, whereas macroR_Freq refers to the synchronization of F0 peaks with prosodic words.

The SpeechRhythmExtractor script [11] used in this paper implements macroR_Freq as a function of the F0 variance within a chunk as in (3) below. While keeping the synchronization interpretation, i.e., larger F0 variation requires a larger number of F0 peaks to convey strong macroR, this implementation becomes fully-automated as it avoids the need to define and count prosodic words.

$$(3)\quad MacroR\_Freq\_f0Var = \frac{number\ of\ F0\ peaks}{F0\ variance}$$

[6] defines macro-rhythm as being the phrase-medial tonal rhythm, and [7] points out that its strength is determined by the number of F0 alternations between peaks and valleys within a phrase. The use of the F0 variance can bring forward a closer look on how the alternations of the peaks and valleys are represented

within a phrase, and wether there is (some) level of regularity for the F0 alternations. By these means, F0 variance showed consistent differences between L2 speech rhythm studies, such as [19].

2.3.3. Speech Rate Measures

Because microR and macroR are affected by speech rate, speech rate is measured to justify the use of normalization measures. Speech rate is computed as the number of syllables per second, and articulation rate as the number of syllables per second, discarding silences. Mean duration and standard deviation of pauses (Pause SD), as well as pause rate, are also computed.

2.4. Statistical Analysis

To assess the sensitivity of the above microR and macroR measures to Japanese-accented English, linear mixed models were computed with R (lme4 package) with each measure as the dependent variable, language (L1 English, L2 English) as the predictor, and participants, TED talk, and chunk as the random factors. To assess the effect of microR on macroR, mixed models were run in R with macroR measures as the dependent variable, microR measures and language as the fixed factors, and participant and TED talk as the random factors.

3. RESULTS

3.1. Pauses, Speech, and Articulation Rates

On the one hand, Japanese speakers closely imitated pause location and duration from the target TED talks (β = -2.78, t = -0.116, p = .90). On the other hand, Japanese speakers inserted additional pauses which led to a significantly higher quantity of pauses (β = 56.15, t = -2.07, p = .03), and, consequently, consistent differences in pause rate between groups (β = 0.01, t = 5.08, p < .001). The speech rate of the imitations was significantly slower than that of the original TED talks (β = -0.50, t = -5.11, p < .001). Moreover, the articulation rate in the imitations was also significantly slower (β = -0.44, t = -6.33, p < .001) as shown in Table 1.

Descriptors	English	Pause SD	Pause rate	Speech rate	Articulation rate
Mean	L1	4.71	0.41	3.83	4.86
	L2	4.12	0.49	3.35	4.46
Median	L1	3.66	0.41	3.79	4.84
	L2	4.09	0.50	3.38	4.46
SD	L1	3.46	0.16	1.05	0.89
	L2	1.96	0.13	0.86	0.64

Table 1. Standard deviation of pauses (Pause SD), Pause rate, Speech Rate, and Articulation Rate in the original talks (L1) and their imitations (L2).

3.2. MacroR and MicroR Measures

Given that L2 speech- and articulation-rate were significantly slower than that of L1, normalized measures, e.g., Varco, nPVI, were analyzed. With regards to macroR, no statistically significant differences between L1 and L2 were obtained for pitch rises (β = 0.03, t = 0.09, p = .92), pitch falls (β = 0.40 t = 1.45 p = 0.14), F0 peak-to-peak distances (β = 0.45 t = 0.34, p = .73), and F0 valley-to-valley distances (β = -0.12, t = -0.07, p = .94) indicating that as a group, Japanese speakers accurately imitated these F0 targets. Consequently, macroR_Var showed no significant differences between L1 and L2 (β = 0.59, t = 0.23, p = .81), as well as macroR_Freq (β = 0.02, t = 0.88, p = .38). However, L2 speakers' macroR_Freq_F0var scores were significantly higher than L1 speakers' (β = 0.06, t = 4.31 p < .001) indicating that Japanese speakers obtained less F0 variability and stronger macroR than those of the original TED samples. The macroR measures from equations macroR_Var, macroR_Freq, and macroR_Freq_F0var are shown in Table 2.

Descriptors	English	MacroR_Var	MacroR_Freq	MacroR_Freq_F0var
Mean	L1	53.50	0.33	0.34
	L2	54.90	0.35	0.40
Median	L1	47.50	0.27	0.33
	L2	52.50	0.29	0.37
SD	L1	35.90	0.29	0.11
	L2	28.60	0.24	0.17

Table 2. MacroR Measures in the original talks (L1) and their imitations (L2).

In contrast to macroR, all microR measures (see Table 3) reached significant differences between groups, showing that Japanese speakers were not as successful at perceiving (and then, imitating) the vowel- and consonant-duration intervals of the original talks. More precisely, %V was significantly higher in the L2 imitation (β < .001, t = 2.07, p = .03) while its counterpart %C was lower (β = -0.01, t = -2.07, p = 0.03) indicating that vowel-intervals in the same chunk occupy more time in L2 than in L1. VarcoV indicates that the duration variation of vowel-intervals in a chunk is larger in the L1 original talk (β = -0.04, t = -2.85, p = .004) while VarcoC shows that for consonant-intervals, it is larger in the L2 imitations (β = 0.04, t = 2.45, p = .01). When duration variation is measured in adjacent intervals in nPVI measures, L2 obtains significantly higher scores than L1 for vowels (β = 0.07, t = 5.97, p < .001), consonants (β = 0.04, t = 4.13, p < .001), vowels or consonants (β < 0.001, t = 7.02, p < .001). This larger

L2 variation between adjacent units is likely related to the frequent inclusion of short epenthetic vowels to break consonant clusters as well as to the slower L2 articulation rate [9, 15, 16].

.	English	%V	%C	VarcoV	VarcoC	nPVI-V	nPVI-C	nPVI-VC
Mean	L1	48.40	51,60	0.68	0.61	0.59	0.60	0.62
	L2	50.00	50,00	0.63	0.65	0.67	0.65	0.68
Median	L1	48.00	52,00	0.65	0.58	0.59	0.60	0.62
	L2	51.00	49,00	0.60	0.63	0.67	0.64	0.68
SD	L1	6.94	6,94	0.16	0.13	0.12	0.10	0.09
	L2	8.15	8,15	0.13	0.15	0.11	0.09	0.08

Table 3. MicroR Measures in the original talks (L1) and their imitations (L2).

3.4. MacroR and MicroR Interactions

Linear mixed models with macroR_Freq_F0var as the dependent variable, language (L2 English and L1 English) and microR variables as fixed effects, and participants and TED talks as random effects were run to explore which microR factors affected macroR. No single fixed effect reached statistical significance, only the following interactions did, namely, language with VarcoV (β = 0.18, t = 1.97, p = 0.04), with nPVI-C (β = 0.35, t = 2.30, p = 0.02), and with nPVI-VC (β = 0.37, t = 2.14, p = 0.03). Figure 1 illustrates the interaction between the nPVI-VC and language showing no correlation for L1 speakers. However, for L2 speakers, reducing nPVI scores towards L1-targets correlates with reducing macroR_Freq_F0var towards L1-targets. Similar effects are observed in the language* nPVI-C interaction. Thus, the higher the nPVI values in L2 – which indicate an effort to increase duration variability at the segmental level to reach L1 duration targets, the higher, and therefore the more regular, macroR_Freq_F0var becomes, missing to match the irregular macroR_Freq L1 index.

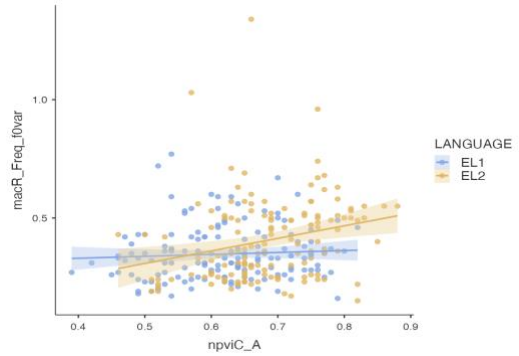

Figure 1. Effect of nPVI-VC on macroR_Freq_F0var by speakers of English as L1 or as L2

4. DISCUSSION AND CONCLUSION

Both speech rate, pauses, and microR results were coherent with previous L2 literature. L2 English imitations of L1 English TED talks were produced at slower speech and articulation rates, and were inserted more pauses than in L1 illustrating the higher cognitive demands of speech planning in L2 e.g., [12]. As for microR, the higher proportion of vowels and smaller variation of vowel-duration intervals in the L2 imitations showed that Japanese speakers transferred to English some of their L1 patterns, e.g., the simpler syllabic structure of Japanese and the lack of duration cues to word prominence. Similar results have been obtained for Japanese speakers of English in previous research, e.g., [13, 14]. The consistently higher L2 nPVI scores, which were also present in [9], were related to the insertion of short epenthetic vowels in complex consonant clusters and to the longer articulation rate in L2 which worked together to increase nPVI scores. Altogether, these microR results answer affirmatively research question 1, namely microR scores captured that microR in Japanese-accented English was not as regular or as strong as in L1 English.

MacroR results to L2 English constitute our main contribution and address research question 2. On the one hand, L1-L2 differences on macroR_Var were non-significant, showing that L2 speakers were highly successful in imitating the L1 pitch contours of the analyzed chunks. On the other hand, L2 macroR-Freq_F0var scores were significantly higher than in L1 due to Japanese speakers' production of F0 peaks that were not present in the original TED talk. These higher values indicated a stronger macroR_Freq index than in L1, answering affirmatively research question 2. Altogether, these macroR_Var and macroR_Freq scores indicate that while Japanese speakers of English produced a sentence F0 contour very similar to that of L1, synchronizing F0 peaks with smaller domains such as prosodic words was still a challenge. In other words, the de-accentuation rate of prosodic words is higher in L1 English.

Finally, research question 3, namely, whether microR interacts with macroR is answered affirmatively so that approximating the larger duration variation of microR in L1 English leads Japanese speakers of English to miss the irregular macroR_Freq of L1 English. Two possible explanations are proposed. The first explanation is based on the deployment of cognitive resources. The higher the effort to imitate vowel and consonant duration intervals, the less resources are directed to process F0 variation causing to miss the irregular macroR of L1 English [12, 15]. In other words, the more cognitive resources are deployed to achieve the strong English microR, the least resources are devoted to achieving the irregular and weak English macroR preventing Japanese speakers from reducing their initial Japanese-like strong macroR.

The second explanation is more linguistically oriented. The learning of the weaker English

macroR_Freq by Japanese speakers is contingent on learning duration cues in words. The hypothesis is that Japanese speakers' deaccentuation of English words will occur more frequently, and therefore, match the weaker English MacroR_Freq, as they use duration to express prominence. To test this hypothesis, future research needs to elucidate what exactly in the vowel interval durations (e.g., English stress, content versus function words) is conditioning macroR_Freq to Japanese learners of English.

7. REFERENCES

[1] Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, *7*(1982), 515-546.

[2] Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for delta C. In Karnowski, P.& Szigeti, I. (eds), *Language and Language Processing: Proceedings of the 38th Linguistics Colloquium, Piliscsaba 2003*. Frankfurt am Main, Germany: Peter Lang Publishing Group, 231-241.

[3] Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*(1-2), 46-63.

[4] Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351-373.

[5] Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), pp. 265-292.

[6] Jun, S. A. (2014). Prosodic typology: By prominence type, word prosody, and macro rhythm. *Prosodic typology II: The phonology of intonation and phrasing*, pp. 520-539.

[7] Prechtel, C. (2020). Quantifying Macro-rhythm in English and Spanish: A Comparison of Tonal Rhythm Strength. University of California, Los Angeles.

[8] Ordin, M., & Polyanskaya, L. (2015). Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. *The Journal of the Acoustical Society of America*, *138*(2), pp. 533-544.

[9] Kisler, T., Reichel, U., Schiel F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, pp. 326-347.

[10] Silva Jr., L. (2022a). VVunitAligner for webMAUS.https://github.com/leonidasjr/VVunitAlignerCode_webMAUS.

[11] Silva Jr., L., Barbosa, P.(2022b). SpeechRhythmExtractor.https://github.com/leonidasjr/SpeechRhythmCode.

[12] Järvinen, K., & Laukkanen, A-M. (2015). Vocal Loading in Speaking a Foreign Language. *Folia Phoniatrica et Logopaedica*, 67, pp. 1-7.

[13] Nagao, J., & Ortega-Llebaria, M. (2022). Micro-and macro-rhythm measures in English and Japanese as first and second languages. In *1st International Conference on Tone and Intonation (TAI). Sonderborg, Denmark*.

[14] Grenon, I. & White, L. (2008). Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In H. Chan, H. Jacob, & E. Kapia, eds., Proceedings of the 32nd Boston Conference on Language Development. Somerville, MA: Cascadilla Press, pp. 155–166.

[15] Krivokapić, J. (2012). Prosodic planning in speech production. In S. Fuchs, M. Weihrich, D. Pape & P. Perrier (Eds.), *Speech planning and dynamics*. (pp. 157-190). Peter Lang.

[16] Igarashi, Y. (2014). Typology of intonational phrasing in Japanese dialects. *Prosodic typology*, *2*, pp. 464-492.

[17] Cole, J., & Shattuck-Hufnagel, S. (2011). The phonology and phonetics of perceived prosody: What do listeners imitate?. In *Twelfth Annual Conference of the International Speech Communication Association*, pp. 969-972.

[18] Petrone, C., D'Alessandro, D., & Falk, S. (2021). Working memory differences in prosodic imitation. *Journal of Phonetics*, *89*(3), pp. 101100. 10.1016/j.wocn.2021.101100

[19] Silva, C., & Arantes, P. (2021). Quantitative analysis of fundamental frequency in Spanish (L2) and Brazilian Portuguese (L1) evidence of learning and language attrition. *Journal of Speech Sciences*, 10, pp. e021003.

[1] Japanese in this paper refers to the Tokyo Japanese variety, i.e., the prestige Japanese variety used in the media and higher education, which is described as a [+accent], [+multiwordAP] by Igarashi in [16].